DOCUMENT RESUME

ED 290 155 CS 211 016

AUTHOR Perkins, Kyle
TITLE A Proposed Research Program for ESL Composition Evaluation.
PUB DATE Nov 86
NOTE 28p.; Paper presented at the Annual Meeting of the National Council of Teachers of English (76th, San Antonio, TX, November 21-26, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *English (Second Language); Error Analysis (Language); Evaluation Criteria; Higher Education; Interrater Reliability; Reading Research; *Research Needs; *Research Projects; Test Reliability; Writing Apprehension; *Writing Evaluation; *Writing Instruction; Writing Research; Writing Skills

ABSTRACT

Based on the premise that composition skills and their evaluation are crucial to the educational process, this paper presents a tentative research program for conducting future English as a second language (ESL) composition evaluation studies. The program developed in the paper covers the following topics as areas which merit further rigorous investigation in ESL composition evaluation: (1) the types of possible reading analyses; (2) the various methods of estimating the reliability of composition ratings; (3) the effects of the different purposes of writing elicited during the testing process; (4) the measurement properties of the rating scales; (5) the estimation of rater effects; (6) the number and the nature of the constructs underlying writing ability; (7) the assessment of errors; (8) the role of writing apprehension in composition assessment; and (9) the need for systematic, replicable analyses of the composing process. After a discussion of the two most commonly used methods of evaluating compositions--holistic scoring and analytical scoring--the paper concludes by noting that although it may be neither feasible nor practical to investigate all of the topics proposed, the results of the proposed studies could augment and enhance an increasing body of ESL composition pedagogy. A list of 35 references and 4 figures conclude the paper. (SKC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED290155

A Proposed Research Program for ESL Composition Evaluation

Kyle Perkins

Southern Illinois University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kyle Perkins

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CS211015

BEST COPY AVAILABLE

Abstract

Today the importance of composition skills cuts across academic disciplines at both the graduate and undergraduate levels of instruction as well as the intensive language training level. Writing ability is complex. The assessment of writing ability is also complex. No assessment procedure is perfectly reliable and valid, and different procedures have not only method effects but are also differentiately sensitive to various aspects of the elicitation and composing processes and to various external factors which are tangentially related to the composing process. Because composition skills and their evaluation are crucial to the educational process, this paper presents a tentative research program for conducting future ESL composition evaluation studies. An attempt has been made to include some of those important factors which are hypothesized to impinge on ESL composition evaluation.

Introduction

The purpose of this paper is to suggest and to motivate a tentative research program for future ESL composition evaluation studies. I believe that ESL composition evaluation practice has now reached a level of sophistication which permits ESL composition researchers to ask themselves where they should next focus their attention and what areas of composition research need further elaboration and enhancement.

I propose to nominate the following topics as areas which merit further rigorous investigation in ESL composition evaluation: (1) the types of possible reading analyses; (2) the various methods of estimating the reliability of composition ratings; (3) the effects of the different purposes of writing elicited during the testing process; (4) the measurement properties of the rating scales; (5) the estimation of rater effects; (6) the number and the nature of the constructs underlying writing ability; (7) the assessment of errors; (8) the role of writing apprehension in composition assessment; and (9) the need for systematic, replicable analyses of the composing process. Obviously, these issues do not constitute a complete list of evaluation issues which are of interest and which are germane to all ESL composition researchers, but this sampler should elicit enough reaction and feedback from composition researchers to give an indication of which issues are the most relevant and the most robust in terms of providing useful information about ESL composition evaluation.

Reading Analyses

Two methods have commonly been used to evaluate compositions: (1) holistic scoring and (2) analytical scoring. Traditionally, holistic scoring has been impressionistic, and it involves one or more readers awarding a single grade based on the total impression of a composition as a whole text or discourse. The impressions are gleaned from an examination of a sample of compositions

for the purpose of determining how many distinct levels of writing are represented in the sample. The number of points on a holistic scale has varied in the literature from four to nine.

Mehrens and Lehmann (1984) suggest the following descriptors for a five-point scale:

1. Superior quality;
2. Above-average quality;
3. Average quality;
4. Below-average quality;
5. Inferior quality (p. 116).

For ESL evaluation perhaps a better set of descriptors is needed with a range of competence from absent to fully developed such as the following:

1. Clearly demonstrates competence;
2. Demonstrates competence;
3. Demonstrates minimal competence;
4. Demonstrates some developing competence;
5. Suggests incompetence;
6. Demonstrates incompetence.

Such a scale is criterion-referenced because the categories describe traits of compositions that differentiate incomplete competence from more completely developed competence. A scoring rubric, complete with a full set of descriptors at each level, is necessary to preclude the standards for each point on the scale from varying from one large reading to the next.

A well-known scoring rubric is the ESL Composition Profile (Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey, 1981). The Profile contains five components which are weighted according to their approximate importance for communication: content (30 points), organization (20 points), vocabulary

(20 points), language use (25 points), and mechanics (5 points). Each component is further broken down into numerical ranges. To use the Profile, a rater must conduct five different holistic evaluations of each composition, one for each of the five components.

Analytical scoring involves the separation of the various features of a composition into components for scoring purposes. One might, for example, elect to use the following analytical scale:

1 2 3 4 5 6

Focus

Mechanics

Organization

Support

The notions that analytical scales can be summed to obtain a composition ability index for individual subjects and that analytical scales permit the identification of a common core element of compositions for instructional purposes are discussed in greater detail in Perkins (1983).

Estimating the Reliability of Composition Evaluations

The reliability of composition evaluations can be defined from at least three different perspectives: (1) the interrater reliability; (2) the check papers analysis; and (3) analysis of variance estimates of reliability and related generalizability components associated with any effect due to rater error.

Reliability has to do with accuracy of measurement. This kind of accuracy is reflected in the obtaining of similar results when measurement is repeated on different occasions or with different instruments or by different raters. This characteristic of reliability is usually referred to as consistency.

Interrater reliability is estimated as the correlation of the evaluations of one rater with those of another. It is assumed that each rater's evaluations are independent. The actual estimated level of reliability will depend on the number of raters.

There are two steps in estimating interrater reliability. First, the evaluations of all raters must be intercorrelated. An average of all correlation coefficients is derived when more than two raters are involved. To derive this average correlation coefficient, one must use a Fisher z transformation because the correlation coefficient is on a non-interval scale. Second, the average correlation coefficient is adjusted using the Spearman-Brown Prophecy Formula (shown below) to make the final reliability estimate reflect the number of raters.

$$r_{tt} = \frac{nr_{A,B}}{1 + (n-1)\, r_{A,B}}$$

r_{tt} = interrater reliability

n = number of raters

$r_{A,B}$ = the average correlation among all raters

The check papers analysis involves having the raters periodically score a set of "check papers" or anchor papers or validation papers during the evaluation process to ascertain that they are continuing to apply the criteria accurately. This procedure implies, of course, that the raters have been introduced to the scoring rubric, that the raters have been given examples of papers that fit within each score level, and that the raters have been given practice in scoring sample essays during a training period.

Interrater reliability, the first method discussed, has been used for years to estimate the reliability of composition evaluations. However, the procedure is flawed for such purposes because it produces inflated estimates due to the fact that only one source of error is recognized--the raters.

Happily, there is a method of estimating reliability that recognizes different sources of error--the analysis of variance estimate of reliability, the third method to be discussed. This model allows one to estimate the variance components for each source of variation and to compute reliability estimates.

The analysis of variance method is a robust model in that it allows a researcher to include a number of factors that affect the reliability of composition evaluations: the choice of topic (Godshalk, Swineford, and Coffman, 1966; Coffman, 1971); different rhetorical modes, e.g., persuasive, narrative, and expository (Kinneavy, 1971; Britton, 1978); the number of writing tasks (Diederich, 1974; Cooper, 1977); the number of topics per mode, and the number of readers per mode (Breland, 1986).

Breland proposed the model displayed as Figure 1 to represent compositions and the various sources of error in composition evaluation.

(Put Figure 1 about here)

The procedure involves a conjunction of the effects of examinees, discourse modes, topics, and raters, and their interactions. Classical test theory estimates of reliability are of a limited use in this situation because, to a large extent, they ignore the multiple sources of variation in the observations.

The researcher can also use Generalizability (G) theory (Brennan, 1983) as an extension of the analysis of variance procedure to take such multiplicity into account in a very explicit manner. In conducting a G-study, the researcher decides in advance which factors have a known or hypothesized influence as one does in a factorial analysis of variance design. The factors may be assumed to be random samples from universes of possible conditions, or they may be fixed, i.e., not randomly sampled from a universe.

In an examinees x topics x discourse modes x raters design, these factors

and their interactions produce variation in the observations. Since these sources of variance and their interactions are identified, the researcher can estimate the reliability of composition evaluations because it is possible to identify the true variation and all sources of error variation and then to compute the ratio of true variance and observed variance.

After the variance components and their contributions have been identified, the researcher can study their effects in a D (decision) study by varying the number of raters, topics, and modes to study the effects on the reliability of the evaluations.

To date there have been no systematic studies of the magnitude of reliability estimation by the check papers, interrater agreement, and the analysis of variance components methods in ESL composition evaluation. While the efficiacy of these methods may interest only those of the psychometric persuasion, the effects of the topics and discourse modes chosen as the test stimuli should be of interest to both the measurement specialist and to the classroom ESL composition teacher.

Effect of the Purposes of Writing Elicited During the Examination

It is apparent from the preceding discussion that the formulation of topics may well have a hypothesized influence on writing performance. It has also been made clear that the variance associated with a specific topic must be distinguished from variance associated with the purpose for a particular writing assignment. Camp (1986) has argued that the content of a topic has a robust effect on writing performance because a topic's content influences the writer's engagement with the topic and the details and evidence which the writer must adduce to develop the paper.

Camp also reported some fascinating issues raised by the holistic scoring of papers written for different purposes. She noted that expository essays (essays that explain, inform, or set forth an idea) were rated more

quickly than either narrative essays (essays that present a series of events to make a point) or persuasive essays (essays that convince the reader of the writer's view). Camp's rationale for these findings was that first language writers have fairly clear and widely shared notions of what expository writing entails. Evaluating the narrative papers was more time consuming because there was less concordance among raters about what constitutes effective narrative prose. Evaluating the persuasive papers was the most time consuming because the raters had to consciously separate the evaluation of the effectiveness of the work samples from their personnal response to the writer's argumentative strategies and proposed solutions.

To date there are no data to indicate whether substantial differential amounts of time are required to evaluate ESL expository, narrative, and persuasive essays. But, if it is found, for example, that trained raters can evaluate 30 expository essays per hour, 20 narrative essays per hour, and only 10 persuasive essays per hour, then the choice of writing task for the essay examination becomes something more than an arbitrary issue in the context of large scale ESL writing assessment.

Measurement Properties of the Writing Scales

The issue of estimating the reliability of composition evaluation has been discussed in a previous section; however, there are certain psychometric properties of the scales per se (both holistic and analytical) that merit investigation.

One issue that comes to mind immediately is the distribution of scores at different proficiency levels (cf. Pyans, 1984). If the researcher has evaluated compositions collected from three different proficiency strata, using, for example, holistic scoring, s/he could calculate the kurtosis, quartile distribution of mean performance, skewness, and standard deviation indices at each proficiency stratum. Kurtosis refers to the peakedness or

flatness of a curve in the center of the distribution. The quartile distribution of mean performance refers to three quartiles which divide a frequency distribution into four equal parts with the first quartile corresponding to the 25th centile, the second quartile to the 50th centile, and the third quartile to the 75th centile. Skewness refers to the symmetry or asymmetry of a distribution curve. Standard deviation is a measure of the scores' deviation from the mean, a measure of score variability. Figure 2 illustrates skewness and kurtosis.

(Put Figure 2 about here)

The components of a holistic scale like the ESL Composition Profile all purportedly measure writing ability. There are at least three implications which follow. First, a matrix of intercorrelations between the components should all be moderately robust and significant. Second, the components should all be represented on the same latent trait. Third, the first two implications should obtain for all proficiency strata.

Given a large enough data set, the composition researcher could submit the component scores from the ESL Composition Profile, for example, or any other suitable analytical scale, to a correlation procedure and an unweighted factor analysis with promax rotation. The purpose of such analyses would be to determine whether the different components are measuring one trait and whether each component contributes some unique information. If one dominant factor is demonstrated for each proficiency level, then one can assume that the components exhibit unidimensionality. In previous ESL composition research there has been little, if any, data offered to demonstrate in an empirical fashion the constructs measured by writing work samples and the scoring procedures used to evaluate them. Research of this nature is fundamental in any measurement enterprise.

The Estimation of Rater Effects

In a previous section I proposed the use of a components of variance model to determine the effects of random variation of the raters on the scores assigned to the writers' work samples. I believe that it is necessary to estimate the raters' effects for the following reasons. First, holistic evaluation can be very subjective because raters may (a) have different standards of severity, (b) distribute scores differently along the scoring continuum, (c) be inconsistent in applying the standards of the evaluation, (d) react to certain elements in the essays, (e) value different aspects of an essay (Jacobs et al., 1981), and (f) interact significantly by topic by mode of discourse and by examinee.

As de Gruijter (1984) notes

> Rater effects can be troublesome when different examinees are judged by different rater teams. For this reason it is important to obtain estimates of rater effects. When these are too large to be neglected, several courses of action are possible: (1) corrections can be made for differences between raters, (2) some of the essays can be rejudged, or (3) rater instructions can be discussed and improved (p. 213).

De Gruijter suggests the use of two simple models for rater effects. One is an additive model which "might be used when no accurate estimates of rater effects are needed, for example, when the results from an analysis are to be used only in a discussion on the impact and origin of differences between raters. When more accurate results are needed, for example, for intended score corrections, ...a nonlinear model might be used" (p. 217).

The formulae for the two models are quite cumbersome and need not be included here. The additive model determines the rater effects to an additive constant, taking into account relative rater leniencies. The non-linear model is an extension of the Rasch one-parameter item response theory model, and it estimates the expected or true score of an examinee taking into consideration the rater and examinee parameters. In essence, the

nonlinear model estimates the true score after the elimination of the rater effect. The potential importance of these models is that they will provide the composition researcher with another tool to isolate and to identify trait variance in a writing examination and remove the nonrandom, nontrait, non-methodological variance associated with the rater.

The Number and the Nature of the Constructs Underlying Writing Ability

In the field of ESL language testing there has been a great debate about whether language proficiency is unidimensional (Oller, 1979; Oller and Perkins, 1978, 1980) or multidimensional (Farhady, 1980; Oller, 1983). The same field of inquiry needs to be transferred to ESL composition evaluation in order to determine whether empirical essay data show evidence for a single factor of a multifactor model of judged writing ability. An ancillary issue is whether this (these) factor(s) relate to objective test measures of writing ability (cf. Rock, 1986).

These questions could be answered by applying confirmatory factor analytic techniques to holistically scored essay data and multiple-choice objective test measures. Some potential multiple-choice tests that might have some usefulness for this project include Test of Standard Written English (TSWE), Test of Ability to Subordinate (TAS), and Revision and Editing Test (RET).

The TSWE is a thirty-minute 50-item multiple-choice test which evaluates an examinee's ability to recognize standard written English. The purpose of TSWE scores is to help place students into appropriate freshman composition courses. It has been administered as a part of the Educational Testing Service's Scholastic Aptitude Test. Two types of questions are included--those on usage and those on sentence correction.

The TAS is a 50-item sentence-combining test which is said to measure specific grammatical points and is recommended for diagnosis and placement

purposes. Kunz (1980) has found that the TAS and cloze scores correlate significantly with the Michigan Test of English Language Proficiency and placement scores.

The RET is an exit exam designed to assess an examinee's ability to revise and to edit. It was developed by Professor Carolyn B. Matalene and her staff at the University of South Carolina. Matalene (1982) described the RET:

> The test is divided into two parts; the first fifteen questions are on units larger than the sentences, the last fifteen questions on how to improve sentences of groups of sentences. Thus revision precedes editing. The test is printed with the entire essay (sentences and paragraphs numbered) in reduced form on page one of the test booklet and stapled so that the student can easily flip back and forth to the essay. Answers are recorded on a computer form to be machine-graded. A perfect score is 150, or thirty times each five-point answer. Passing should theoretically be 120, or thirty times each four-point answer, since both the four- and the five-point answers are correct according to departmental standards. The questions ask students to discover thesis and topic sentences; to make judgments about level of language, voice, coherence, development, logic, diction; to discern methods of development, errors of logic, unstated assumptions, sentence variety, patterns of errors, effectiveness of examples; and to offer suggestions for revision at every level. Questions about sentences ask how best to combine or revise or punctuate or correct or reword with alternative choices of revised sentences to choose from. In short, students are asked how best to revise and edit the entire essay. Because of the test format, questions sometimes ask what error or problem is the most serious; in this regard the bias of the test is that structural errors are more important than surface ones, that editing is more important than proofreading (p. 373).

To conduct the factor analytic study the ESL composition researcher could pose and test a series of factor models using a LISREL program (Joreskog and Sorbom, 1985) for the maximum likelihood estimation of the factor models. This program allows the estimation process to begin with simple models of writing proficiency, i.e., a single factor, and to proceed to more complex models, i.e., three or four factors. Various goodness to fit measures could be used to determine how well the data fit the model. Of particular interest would be the loadings of various objective test measures of writing ability. Studies of this type could shed light on whether there is a

dominant general ESL writing proficiency factor with definable subfactors and whether holistic evaluations of ESL writing performance on familiar versus unfamiliar topics had an effect on the relationships with objective multiple-choice tests of writing (cf. Rock, 1986).

The Assessment of Errors

Hull (1984) has noted that an increased interest in discourse features and supra-sentential features has come about at the expense of interest in the validity and reliability of the assessment of sentence-level errors. A second reason offered by Hull for the lack of interest in error assessment is the widespread assumption that sentence-level error assessment is easy, and the bottom up process consisting of sentence-level decisions has been largely ignored in the composing process.

Hull and a colleague have attempted to characterize the processes used by raters to identify and classify sentence-level errors in written texts. Their findings indicate that (1) some errors such as spelling errors are easily recognized and are easily categorized, (2) some errors are easy to identify but difficult to categorize, (3) some errors were not errors but reflected judgments of style and taste, and (4) raters tend to categorize errors by imagining the writer's intentions and by reconstructing the writer's intended meaning. A second language acquisition researcher conducting an error analysis immediately recognizes these problems because of the difficulties of deciding whether a particular error should be categorized as a developmental, a language transfer, or a performance error. Hull has suggested systematic research on error patterns, the creation of better, meaningful taxonomies, and new tasks for testing the assessment of sentence-level performance.

I applaud and commend Hull's efforts for such research is also

desperately needed in second language acquisition research and in ESL composition research. A rationale and a justification for further research on the assessment of errors can, I think, be found in Krashen's (1982) discussion of what can be monitored (pp. 92-94). In his discussion of the successful Monitor performer Krashen notes that the number of rules actually learned by the best students is very limited (see Figure 3). Our study of successful Monitor use would be greatly enhanced by a new paradigm of error assessment proposed by Hull.

(Put Figure 3 about here)

The Role of Writing Apprehension in Composition Assessment

Writing apprehension, or writing test anxiety, is a construct that has been investigated systematically during the last decade. John Daly and his colleagues at the University of Texas have been the leading first language researchers in writing apprehension and its effects on writing assessment. Daly and his associates developed an instrument, the Daly-Miller Writing Apprehension Test (1975), to detect writing apprehension, and the constructs of "approach to and avoidance of" writing are bases for the instrument.

Leah Miller (1984) conducted a rigorous systematic analysis of writing apprehension among ESL composition students. She used an adaption of the Daly-Miller Writing Apprehension Test and the Thompson Writing Attitude Survey (1979) as an elicitation instrument. Miller's subject pool consisted of foreign students enrolled in the first of a two-semester sequence of freshman composition. She reported the following:

> The Writing Attitude Survey (Miller's instrument) composed of items from the Writing Apprehension Test and the Thompson Writing Attitude Survey identifies high, normal, and low levels of writing anxiety in EFL student writers. High anxiety is not a personality characteristic of any of the native language groups in this subject pool. The individual teachers did not induce high levels of anxiety in their students, but they may have reduced anxiety levels over the semester. In addition, high levels of

> anxiety as determined by the Writing Attitude Survey do not predict that individuals will perform poorly in writing classes. Individuals may suffer from the debilitating effects of anxiety and, as a result, they may dislike composing and composition classes, but the effects of anxiety do not portend failure in these classes (p. 47).

Although Miller did not find evidence that writing apprehension had predictive validity, it would be interesting to investigate the extent to which writing apprehension is found at different levels of English language proficiency, say TOEFL 351-400, 401-450, 451-500, 501-550, etc., and whether there would be a significant difference in levels of writing apprehension across the proficiency strata.

Writing apprehension could be a source of nonrandom, nontrait variance. Since measures of writing apprehension exist, one could include writing apprehension as a source of variance into the analysis of components of variance model presented in Figure 1. Inclusion of writing apprehension in the model would permit a study of its interaction with the choice of topic, different rhetorical modes, and the number of topics per mode. Such a model would accommodate yet a still better picture of the relationship between trait variance and observed variance in the evaluation of ESL compositions.

The Need for Systematic, Replicable Analyses of the Composing Process

Within the last decade much has been written about the composing process. Various models have been offered which depict the process per se. The simplest model is a linear chronological model: Prewriting Writing Rewriting Editing. A more elaborate model which depicts composing and revising is the one provided by Smit (1986) and shown in Figure 4.

(Put Figure 4 about here)

Models of the composing process are heuristic; they provide a plan for characterizing the composing processes of ESL composition students. Pianko (1977) and Perl (1979) have provided a usable paradigm for studying the ESL composing process. Perl coded the composing behaviors exhibited by

her students during the composing process and distinguished the following operations from the major categorical divisions of reading, talking, and writing:

(1) General planning-organizing one's thoughts for writing, discussing how one will proceed.
(2) Local planning-talking out what idea will come next.
(3) Global planning-discussing changes in drafts.
(4) Commenting-sighing, making a comment or judgment about one's topic.
(5) Interpreting-rephrasing the topic to get a "handle" on it.
(6) Assessing-making a judgment about one's writing, may be positive or negative.
(7) Questioning-asking a question.
(8) Talking leading to writing-voicing ideas on the topic, tentatively finding one's way, but not necessarily being committed to or using all one is saying.
(9) Talking and writing at the same time-composing aloud in such a way that what one is saying is actually being written at the same time.
(10) Repeating-repeating written or unwritten phrases a number of times.
(11) Reading related to the topic:
    (a) Reading the directions
    (b) Reading the question
    (c) Reading the statement
(12) Reading related to one's own written product:
    (a) Reading one sentence or a few words
    (b) Reading a number of sentences together
    (c) Reading the entire draft through
(13) Writing silently.
(14) Writing aloud.
(15) Editing
    (a) adding syntactic markers, words, phrases, or clauses
    (b) deleting syntactic markers, words, phrases, or clauses
    (c) indicating concern for a grammatical rule
    (d) adding, deleting, or considering the use of punctuation
    (e) considering or changing spelling
    (f) changing the sentence structure through embedding, coordination or subordination
    (g) indicating concern for appropriate vocabulary (word choice)
    (h) considering or changing verb form
(16) Periods of silence (pages 320-321).

Perl's system allows the composition researcher to categorize and to analyze the writers' observable behaviors--a procedure that is capable of being replicated from one setting to another. The categorized data can be recorded on to composing style sheets for each student writer to show how, not what they wrote. The charts allow the researcher to determine the

following information:

(1) the amount of time spent during prewriting;
(2) the strategies used during prewriting;
(3) the amount of time spent writing each sentence;
(4) the behaviors that occur while each sentence is being written;
(5) when sentences are written in groups or "chunks" (fluent writing);
(6) when sentences are written in isolation (choppy or sporadic writing);
(7) the amount of time spent between sentences;
(8) the behaviors that occur between sentences;
(9) when editing occurs (during the writing of sentences, between sentences, in the time between drafts);
(10) the frequency of editing behavior;
(11) the nature of the editing operations; and
(12) where and in what frequency pauses or periods of silence occur in the process (p. 322).

If the ESL composing process is to be studied in a rigorous, systematic fashion, Perl's paradigm allows the ESL composition researcher to study any sequential patterns that occur between and within students and writing sessions. To date we have no idea as to whether ESL students' writing behaviors are random or systematic, recursive or linear. The results from this kind of study would surely be of interest to second language acquisition researchers and to those keen on language teaching pedagogy.

An Afterthought

If ESL composition teachers accept the validity of the composing process, then ESL composition testing behavior must change. From the preceding discussion of Perl's research paradigm, and from our knowledge of the composing process, we can infer that we attempt to inculcate one behavior in the ESL composition class and test yet another behavior in the testing hall. The composing process entails that students must have time to think aloud about their topic, to gather information, and to assemble the data into a format which will allow a thesis and a logical method of organization to evolve. Then a draft is written and submitted to the instructor for guided revision. After one or more revised drafts, the students are ready

to begin editing the final draft. But all too often ESL students face the following circumstances during the final proficiency tests of writing ability: "Choose one of the following topics and write 250-300 words of well-organized prose in 50 minutes."

If we accept the validity of the composing process, this testing procedure is patently invalid: we are teaching X and testing Y. If the composing process is taught in the classroom, then attempts must be made to replicate the composing process in the testing hall--a change which carries many implications such as extending the writing test period or possibly giving the students a set of material to read and to react to.

If we can accommodate and replicate the composing process in the testing hall, we would thus afford the examinees the opportunity to transform their early drafts, which can be labelled Writer-Based prose, to more refined, subsequent drafts, which can be called Reader-Based prose. Flower (1981) distinguished Writer- and Reader-Based prose:

> In function, Writer-Based prose is a verbal expression written by a writer to himself and for himself. It is the record and the working of his own verbal thought. In its structure, Writer-Based prose reflects the associative, narrative path of the writer's own confrontation with her subject. In its language, it reveals her use of privately loaded terms and shifting but unexpected contexts for her statements. In contrast, Reader-Based prose is a deliberate attempt to communicate something to a reader. To do that, it creates a shared language and shared context between writer and reader. It also offers the reader an issue-centered rhetorical structure rather than a replay of the writer's discovery process. In its language and structure, Reader-Based prose reflects the purpose of the writer's thought; Writer-Based prose tends to reflect its process. Good writing, therefore, is often a cognitively demanding transformation of the natural but private expressions of Writer-Based thought into a structure and style adapted to a reader (p. 269).

If the examinees are not allowed sufficient time to reflect on the topic and to plan the content and the logical organization, they may

experience a cognitive overload because they are attempting to attend to too many factors simultaneously. Bereiter and Scardamalia (1983) have discussed this phenomenon observed in children first learning to write. They point out that writers frequently encounter a capacity limit problem and become engrossed with the mechanical aspects of the composing process, leaving little or no time for concerns with content. Flower and Hayes (1980) offer a concise description of the factors that can cause a cognitive overload during the composing process:

> A writer in the act of discovery is hard at work searching memory, forming concepts, and forging a new structure of ideas, while at the same time trying to juggle all the constraints imposed by his or her purpose, audience, and language itself. Discovery, the event, and its product and new insights are only the end result of a complicated intellectual process (p. 21).

Conclusion

It may be neither feasible nor practical to investigate all of the topics proposed in this paper. Still, one of the purposes of composition evaluation is to provide a washback effect on pedagogy. The results of many of these proposed studies could augment and enhance an ever-increasing body of ESL composition pedagogy in addition to providing valuable information about ESL composition students and the work samples which they produce for evaluation. Given these considerations, I believe the research program proposed in this paper is a legitimate one.

References

Bereiter, C., & Scardamalia, M. (1983). Does learning to write have to be so difficult? In A. Freedman, I. Pringle & J. Yalden (Eds.), Learning to write: First language/second language (pp. 20-33). London: Longman.

Breland, H. (1986). Study objectives and outcomes of reliability and predictive validity analyses. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April.

Brennan, R. (1983). Elements of generalizability theory. Iowa City: American College Testing Program.

Britton, J. (1978). The composing process and the functions of writing. In C. Cooper & L. Odell (Eds.), Research on composing: Points of departure (pp. 13-28). Urbana: National Council of Teachers of English.

Camp, R. (1986). The writing tasks and their scoring. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April.

Coffman, W. (1971). Essay examinations. In R. L. Thorndike (Ed.), Educational measurement. 2nd ed. Washington, D.C.: American Council on Education.

Cooper, C. (1977). Holistic evaluation of writing. In C. Cooper & L. Odell (Eds.), Evaluating writing: describing, measuring, judging (pp. 3-31). Urbana: National Council of Teachers of English.

Daly, J. & Miller, M. (1975). The empirical development of an instrument to measure writing apprehension. Research in the Teaching of English, 9, 242-249.

De Gruijter, D. (1984). Two simple models for rater effects. Applied Psychological Measurement, 8, 213-218.

Diederich, P. (1974). Measuring growth in English. Urbana: National Council of Teachers of English.

Farhady, H. (1980). Justification, development, and validation of functional language testing. Ph.D. dissertation, University of California, Los Angeles.

Flower, L. (1980). Writer-based prose: a cognitive basis for problems in writing. In G. Tate & E. Corbett (Eds.), The writing teacher's source-book (pp. 268-292). New York: Oxford University Press.

Flower, L. & Hayes, J. (1980). The cognition of discovery: defining a rhetorical problem. College Composition and Communication, 31, 21-32.

Fyans, L. (1984). The psychometric evaluation of the analytic writing scale. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April.

Godshalk, F., Swineford, F., & Coffman, W. (1966). The measurement of writing ability. Research Monograph No. 6. New York: College Entrance Examination Board.

Hull, G. (1984). Assessing errors in writing. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, F., & Hughey, J. (1981). Testing ESL composition. Rowley, MA: Newbury House.

Joreskog, K., & Sorbom, D. (1985). Lisrel 6, Analysis of linear structural relationships by the method of maximum likelihood: User's guide. Mooresville, IN: Scientific Software, Inc.

Kinneavy, J. (1971). A theory of discourse. Englewood Cliffs, N.J.: Prentice-Hall.

Krashen, S. (1982). Principles and practice in second language acquisition. Oxford: Pergamon Press.

Kunz, L. (1980). English as a second language. In A. Trillin and Associates (Eds.), Teaching basic skills in college (pp. 145-207). San Francisco: Jossey-Bass.

Matalene, C. (1982). Objective testing: politics, problems, possibilities. College English, 44, 368-381.

Mehrens, W. & Lehmann, I. (1984). Measurement and evaluation in education and psychology. 3rd. ed. New York: Holt, Rinehart and Winston.

Miller, L. (1984). Evaluating writing anxiety in EFL students. M.A. thesis, Southern Illinois University, Department of Linguistics.

Oller, J. (1979). Language tests at school. London: Longman.

Oller, J. (1983). Issues in language testing research. Rowley, MA: Newbury House.

Oller, J., & Perkins, K. (1978). Language in education: testing the tests. Rowley, MA: Newbury House.

Oller, J., & Perkins, K. (1980). Research in language testing. Rowley, MA: Newbury House.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. TESOL Quarterly, 17, 651-671.

Pianko, S. (1977). The composing acts of college freshman writers. Ed.D. dissertation, Rutgers University.

Perl, S. (1979). The composing processes of unskilled college writers. Research in the Teaching of English, 13, 317-326.

Rock, D. (1986). Essay tests: what is being measured, how does it differ from what is being measured by multiple-choice tests. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April.

Roscoe, J. (1975). Fundamental research statistics for the behavioral sciences. 2nd ed. New York: Holt, Rinehart and Winston.

Smit, D. (1986). On being clear: the implications of reading research for the teaching of writing. Unpublished paper, University of Iowa, Department of English.

Thompson, M. (1979). Writing anxiety and freshman composition. Urbana: Clearinghouse on Reading and Communication Skills, National Council of Teachers of English.

Figure 1
Analysis of Components of Variance

$$Y_{emtr} = m + a_e + b_m + c_{t(m)} + d_{r(m)} + ab_{em} + ac_{et(m)} + ad_{er(m)} + cd_{tr(m)} + e_{emtr}$$

where:

Y_{emtr} = response of examinee e to topic t in mode m when read by rater r

m = mean response across all examinees

a_e = random variation associated with examinee e

b_m = fixed deviation associated with discourse mode m

c_t = random variation associated with topic t (within mode m)

$d_{r(m)}$ = random variation associated with rater r (within mode m)

ab_{em} = random variation associated with examinee by mode interaction

$ac_{et(m)}$ = random variation associated with examinee by topic interaction (within mode)

$ad_{er(m)}$ = random variation associated with examinee by reader interaction (within mode)

$cd_{tr(m)}$ = random variation associated with topic by rater interaction (within mode)

e_{emtr} = error variance (junk, unexplained, non-trait, non-methodological variance)

Estimate of Reliability

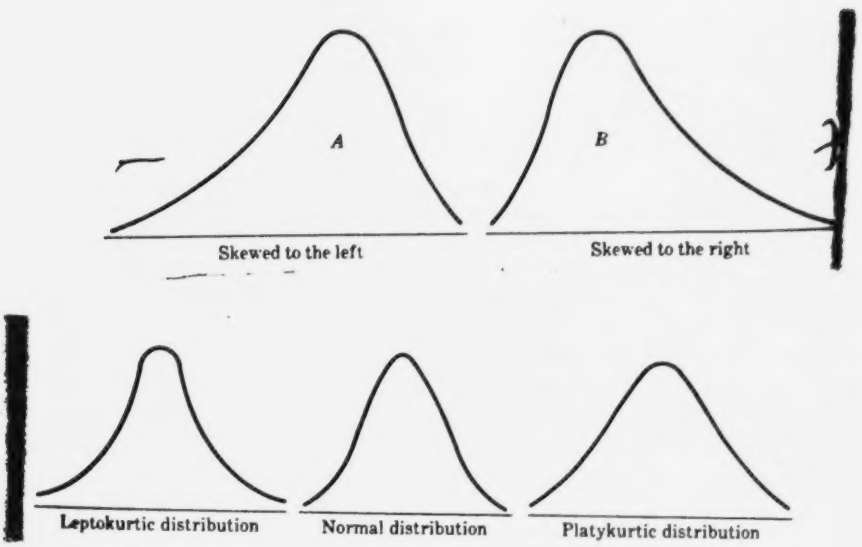
$$r_{tt} = \frac{s_a^2 + \frac{s_{ab}^2}{M}}{s_a^2 + \frac{s_{ab}^2}{M} + \frac{s_c^2 + s_{ac}^2}{MT} + \frac{s_d^2 + s_{ad}^2}{MR} + \frac{s_{cd}^2 + s_e^2}{MTR}}$$

M = number of modes

T = number of topics per mode

R = number of raters per mode

Figure 2
Skewness and kurtosis



(Roscoe, 1975, pp. 46-47)

Figure 3
Monitor Use

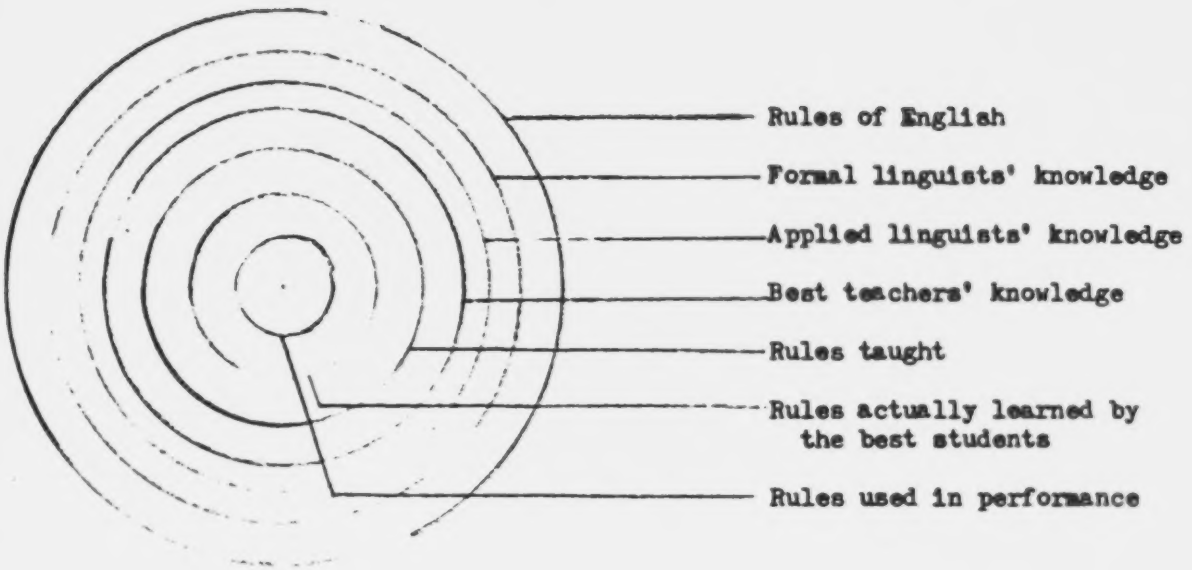
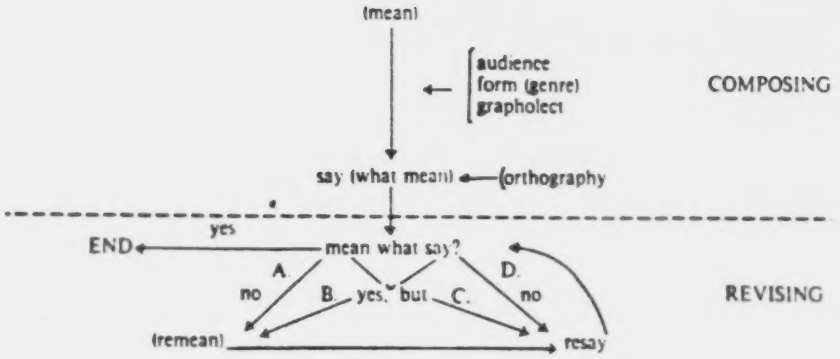


Figure 4
The Composing Process



(Smit, 1986, p. 17)